Single Nucleotide Polymorphisms in *Mycobacterium tuberculosis* Structural Genes—Response to Dr. Musser

To the Editor: In his letter on single nucleotide polymorphisms in *Mycobacterium tuberculosis*, Dr. Musser indicates that genome strain CDC1551 has not been published. Cole et al. (1) described some of the biology of *M. tuberculosis* based on the genome sequence data. The actual sequence, while not published, is in GenBank (Accession NC00962), the sequence data are available at www.sanger.ac.uk, and the annotation is available at http://genolist.pasteur.fr/TubercuList/ . We have a manuscript in preparation using a method of whole genome comparison (2) to evaluate the sequence diversity of strains H37Rv and CDC1551 and applying the information to the analysis of >150 clinical isolates. The complete sequence data and annotation for strain CDC1551 have been available for over a year at www.tigr.org and www.tigr.org/CMR, and periodic updates are provided. In addition, we are preparing to submit the strain CDC1551 sequence and annotation to GenBank (Accession AE000516).

We agree that sequencing accuracy in assessing comparative single nucleotide polymorphism (SNP) data is important. The error frequency suggested by Dr. Weinstock ("Error frequency in a finished sequence has never been precisely measured but is thought to be one error [frameshift or base substitution] in 10^3 to 10^5 bases" [3]) is not supported by any evidence. The whole-genome shotgun sequencing method developed by The Institute for Genomic Research (TIGR) (4) and adopted by many others is highly accurate because of the following qualities: 1) high redundancy in shotgun sequencing (average 7.9-fold for the strain CDC1551 project with a minimum of 2-fold coverage for any nucleotide); 2) assignment of quality values to each nucleotide base; 3) adoption of assembly programs that use quality values for consensus building; and 4) manual editing of electropherograms as necessary.

These methods were applied to the *M. tuberculosis* genome sequencing project. In comparing the CDC1551 and H37Rv strains, it is reasonable to suspect that the SNPs also have the potential to be results of sequencing errors. The sequence differences were verified by two independent methods. One hundred SNPs were chosen at random, and the base calls were independently verified by inspection of the original electropherograms at TIGR (CDC1551) and the Sanger Center (H37Rv). A second method, independent of sequencing, was also used to confirm the base calls of these 100 SNPs. The visual inspection of the electropherograms and the sequencing independent method were in good agreement and indicated that 80 (91%) of 88 successful assays of the nucleotide differences were genuine.

Since our initial report, we have improved our methods for overlaying the annotation of open reading frame coordinates onto our analysis of the coordinates of nucleotide substitutions. Approximately 7% of the genome is noncoding, and approximately 15% of the substitutions are in these regions.

Dr. Musser is correct in pointing out that the substitution frequency expressed in Fraser et al. (5), based on our preliminary annotation of our *M. tuberculosis* sequence data, is not an equivalent comparison to the synonymous substitution frequency derived by his method of sequencing a select set of genes over a wide range of *M. tuberculosis* strains. He uses the methods of Li et al. (6), among the most widely accepted, for the calculation of nucleotide substitution frequencies and derives a D_s value of <0.01 synonymous substitutions per 100 synonymous sites. Our preliminary data presented the frequency of total nucleotide substitutions at all positions (coding [synonymous and nonsynonymous] and noncoding) of the two recently sequenced strains, H37Rv and CDC1551. Our manuscript in preparation comparing the two *M. tuberculosis* strains will contain an analysis of synonymous substitutions. However, while Dr. Musser compared a select group of genes over perhaps several hundred strains, our frequency will be based on a genome-wide comparison between two strains.

Robert Fleischmann
The Institute for Genomic Research
Rockville, Maryland, USA

References

1. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393:537-44.
2. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res 1999;27:2369-76.
3. Weinstock GM. Genomics and bacterial pathogenesis. Emerg Infect Dis 2000;6:496-504.
4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 1995;269:496-512.
5. Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S. Comparative genomics and understanding of microbial biology. Emerg Infect Dis 2000;6:505-12.
6. Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2000;2:150-512.